Mathmatial Foundations of Reinforcement Learning

Markov decision processes: basics

Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

Fall 2023

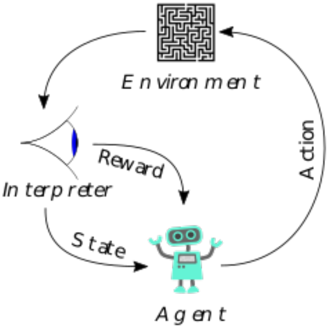
Outline

Markov decision processes

Policy evaluation

Reinforcement learning

In reinforcement learning (RL), an agent learns through interaction with the (unknown) environment.



RL has deep connection with control theory, and is also sometimes called approximate dynamic programming. It can be viewed as a type of optimal control theory with no pre-defined model of the environment.

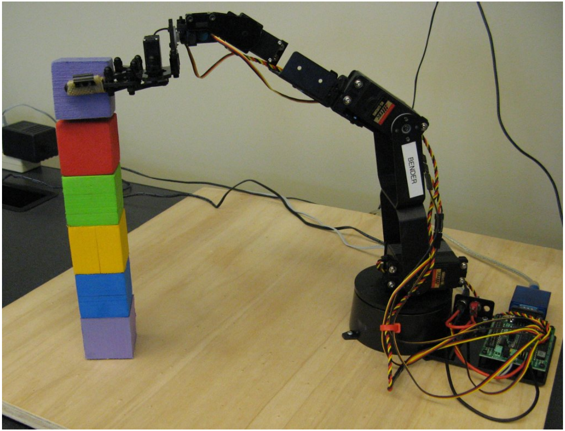
Applications of RL

RL can be applied to many different areas.

- Robotics: in which direction and how fast should a robot arm move?
- Mobility: where should taxis go to pick up passengers?
- Transportation: when should traffic lights turn green?
- Recommendations: which news stories will users click on?
- Network configuration: which parameter settings lead to the best allocation of resources?

Similar to multi-armed bandits, but with a notion of state or context.

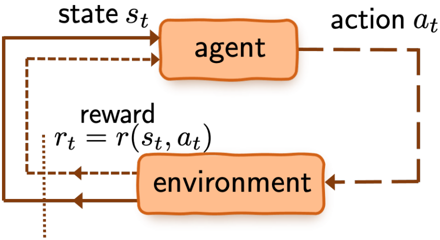
Example: grasping an object



RL reinforces the agents' decisions over time by observing the reward and state that result from taking different actions.

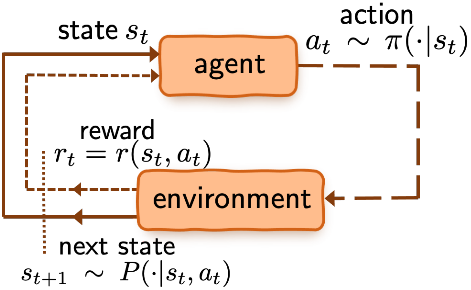
Markov decision processes

Infinite-horizon Markov decision process (MDP)



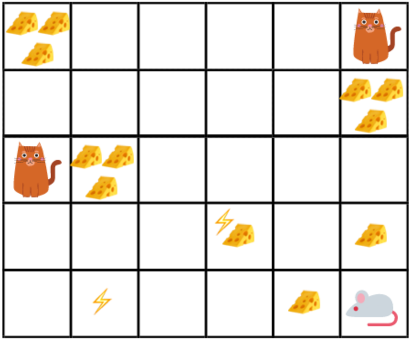
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward

Infinite-horizon Markov decision process (MDP)



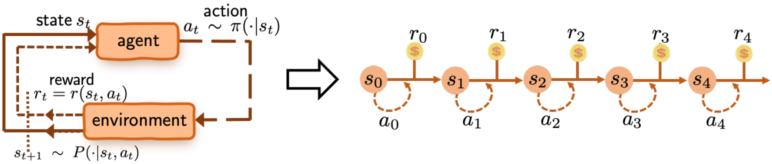
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule), deterministic or random
- $P(\cdot|s, a)$: transition probabilities

Help the mouse!



- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese $(+1)$, electricity shocks (-1), cats (-10000)
- policy $\pi(\cdot|s)$: the way to find cheese

Value function

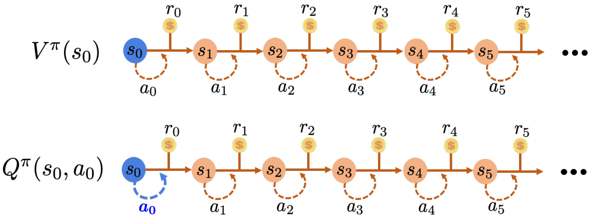


Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

- $(a_0, s_1, a_1, s_2, a_2, \cdots)$: induced by policy π
- $\gamma \in [0, 1)$: discount factor,
  - γ close to 0 leads to "myopic" evaluation
  - γ close to 1 leads to "far-sighted" evaluation

Q-function (action-value function)


$V^\pi(s_0)$
r_0 r_1 r_2 r_3 r_4 r_5
s_0 s_1 s_2 s_3 s_4 s_5
a_0 a_1 a_2 a_3 a_4 a_5
$Q^\pi(s_0, a_0)$
r_0 r_1 r_2 r_3 r_4 r_5
s_0 s_1 s_2 s_3 s_4 s_5
$\boldsymbol{a_0}$ a_1 a_2 a_3 a_4 a_5

Q-function of policy π:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

- $(\cancel{a_0,}\ s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Effective horizon

Since $r(s,a) \in [0,1]$,

$$0 \leq V^\pi(s), Q^\pi(s,a) \leq \frac{1}{1-\gamma}.$$

Often think of $\frac{1}{1-\gamma}$ as the **effective horizon** of the problem.

Why Markov transitions?

- By the Markovian property,

$$\begin{aligned} P(s_{t+1}, s_t, \ldots, s_0) &= P(s_0)P(s_1|s_0)P(s_2|s_1, s_0)\ldots P(s_{t+1}|s_t, \ldots, s_0) \\ &= P(s_0)P(s_1|s_0)P(s_2|s_1, \cancel{s_0})\ldots P(s_{t+1}|s_t, \cancel{\ldots, s_0}) \\ &= P(s_0)\prod_{i=0}^{t} P(s_{i+1}|s_i). \end{aligned}$$

  Low computation and memory complexity!

- The world is Markovian when the state space is large enough. For example, if $s_{t+1} \sim P(\cdot|s_t, s_{t-1})$ depends on the previous two steps, by working with $\widetilde{s}_t = (s_t, s_{t-1})$ (and $s_{-1} = s_0$), we have

$$\widetilde{s}_{t+1} \sim P(\cdot|\widetilde{s}_t)$$

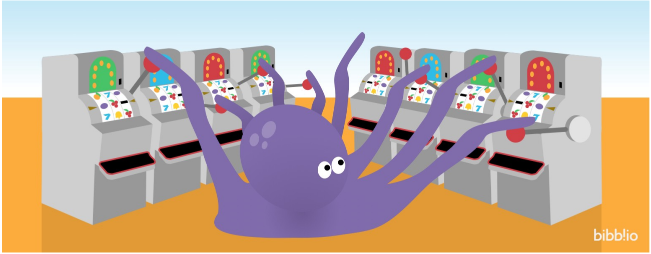
  is Markovian.

All models are wrong, but some are useful

Why discounting?

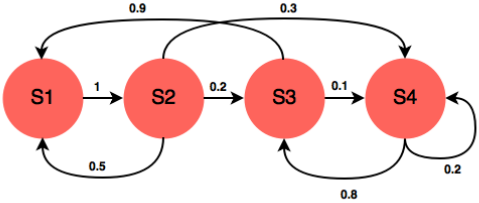
- Mathematically convenient: the limit always exists
- Immediate rewards earn more interest than future rewards
- Account for variability and uncertainty in the future which may not be fully captured
- Undiscounted MDP is possible, e.g. if all sequences terminate (like in a maze or game).
- Alternatives: average reward and finite-horizon episodic settings.

Reduction to multi-arm bandits



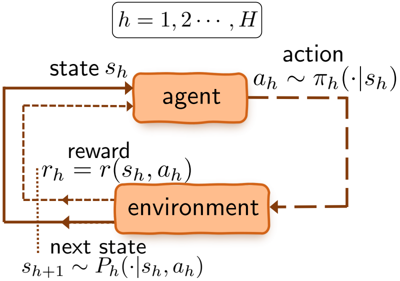
- No state transition: $\mathcal{S}$ is a singleton
- The reward function is action-dependent (action = arm): $r(a)$
- Short-horizon planning: discount factor $\gamma = 0$
- The value of a policy π becomes $V^{\pi} := \mathbb{E}_{\pi}[r(a)]$.

Reduction to Markov reward process



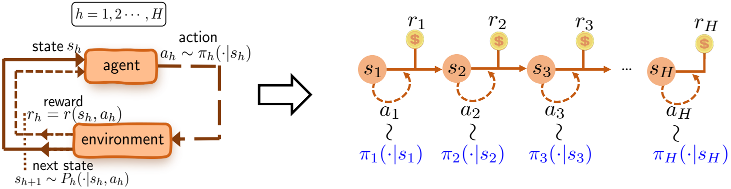
- No action selection: $\mathcal{A}$ is a singleton
- The transition kernel defines a Markov chain
- The reward function is state-dependent: $r(s)$
- The value becomes $V(s) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t) \,|s_0 = s]$.

Finite-horizon episodic MDP



- H: horizon length
- $\mathcal{S}$: state space with size S
- $\mathcal{A}$: action space with size A
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{h=1}^{H}$: policy (or action selection rule)
- $P_h(\cdot \mid s, a)$: transition probabilities in step h

Value function and Q-function



$$V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s\right]$$

$$Q_h^\pi(s, a) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s, a_h = a\right]$$

- execute policy π to generate sample trajectory

Basic tasks

Policy evaluation:

- given a policy π, how good is it?

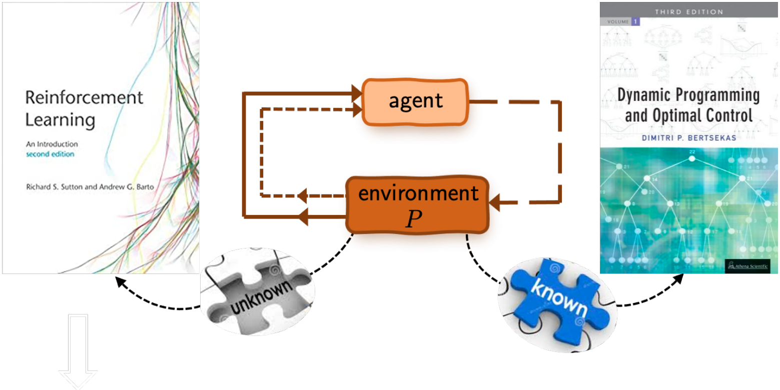
Policy improvements:

- given a policy π, can we find a better one?

Policy optimization:

- can we find the best policy for the given MDP?

Planning versus learning



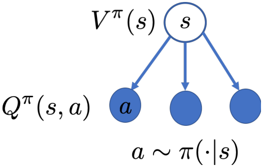
- **Planning:** solve for a desired policy given model specification
- **Learning:** learn a desired policy from samples w/o model specification

  We'll focus on planning first.

Policy evaluation

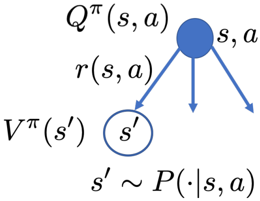
Policy evaluation: evaluating V via Q

$$
\begin{aligned}
V^\pi(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right] \\
&= \sum_{a \in \mathcal{A}} \pi(a_0 = a | s = s_0) \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a\right]}_{=: Q^\pi(s,a)} \\
&= \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s, a)]
\end{aligned}
$$



Policy evaluation: evaluating Q via V

$$
\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}\Big[r(s,a) + \sum_{t=1}^{\infty} \gamma^t r(s_t,a_t) \,\Big|\, s_0 = s, a_0 = a\Big] \\
&= \mathbb{E}\left[r(s,a)\right] + \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r(s_t,a_t) \,\Big|\, s_1 = s', s_0 = s, a_0 = a\Big]\right] \\
&= \mathbb{E}\left[r(s,a)\right] + \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r(s_t,a_t) \,\Big|\, s_1 = s'\Big]\right] \\
&= \mathbb{E}\left[r(s,a)\right] + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\underbrace{\left[\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \,\Big|\, s_0 = s'\Big]\right]}_{=:V^\pi(s')} \\
&= \mathbb{E}\left[r(s,a)\right] + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V^\pi(s')\right]
\end{aligned}
$$



Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{\mathbb{E}[r(s,a)]}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

The value/Q function can be decomposed into two parts:

- immediate reward $\mathbb{E}\left[r(s,a)\right]$
- discounted value of at the successor state $\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V(s')$

Richard Bellman

Matrix-vector representation

- Plugging Q^π into V^π, we have

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)] + \gamma \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s), s'\sim P(\cdot|s,a)}[V^\pi(s')].$$

- Let P^π be the state-state transition matrix induced by π, namely,

$$P^\pi(s'|s) = \sum_{a\in\mathcal{A}} \pi(a|s)P(s'|s,a).$$

- We can write the above in a matrix-vector form as

$$V^\pi = r^\pi + \gamma P^\pi V^\pi,$$

where $V^\pi = [V^\pi(s)]_{s\in\mathcal{S}}$, and $r = \left[\mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]\right]_{s\in\mathcal{S}}$.

a similar treatment applies to Q^π, too.

Solving the Bellman's consistency equation

$$V^\pi = r^\pi + \gamma P^\pi V^\pi \quad \implies \quad V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$$

Invertibility of $I - \gamma P^\pi$: Gershgorin's circle theorem, or for any $x \in \mathbb{R}^{|\mathcal{S}|}$, verify

$$\begin{aligned} \|(I - \gamma P^\pi)x\|_\infty &\geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty \\ &\geq \|x\|_\infty - \gamma \|x\|_\infty \qquad (\|P^\pi x\|_\infty \leq \|P^\pi\|_1 \|x\|_\infty = \|x\|_\infty) \\ &\geq (1-\gamma)\|x\|_\infty \\ &> 0. \end{aligned}$$

Thus, $I - \gamma P^\pi$ is full rank and invertible.

Computationally expensive for problems with large state space!

Bellman's policy operator

Bellman's policy operator: denote the operator $\mathcal{T}^\pi : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ as

$$\forall V \in \mathbb{R}^{|\mathcal{S}|} : \qquad \mathcal{T}^\pi(V) = r^\pi + \gamma P^\pi V.$$

Fixed-point equation:

$$V = \mathcal{T}^\pi(V)$$

- V^π is the unique fixed point of this fixed-point equation.

Contraction property of the Bellman's operator

Lemma 1

The operator $\mathcal{T}^\pi$ is a γ-contraction on $\mathbb{R}^{|\mathcal{S}|}$, i.e. for any V and V' in $\mathbb{R}^{|\mathcal{S}|}$, it follows

$$\|\mathcal{T}^\pi(V) - \mathcal{T}^\pi(V')\|_\infty \leq \gamma \|V - V'\|_\infty .$$

Proof: For any V and V',

$$\begin{aligned}\|\mathcal{T}^\pi(V) - \mathcal{T}^\pi(V')\|_\infty &= \|\gamma P^\pi V - \gamma P^\pi V'\|_\infty \\ &\leq \gamma \|P^\pi\|_1 \|V - V'\|_\infty = \gamma \|V - V'\|_\infty ,\end{aligned}$$

using $\|P^\pi\|_1 = 1$.

Fast computation without inversion

Value iteration for policy evaluation

For $t = 0, 1, 2, \ldots$

$$V^{(t+1)} = \mathcal{T}^\pi(V^{(t)}).$$

Linear convergence:

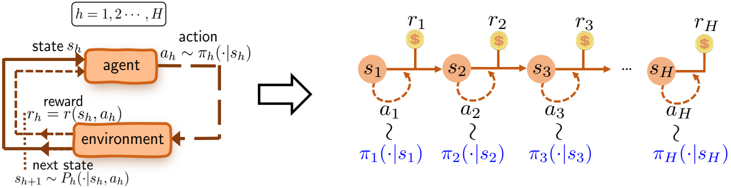
$$\begin{aligned}
\|V^{(t+1)} - V^\pi\|_\infty &= \|\mathcal{T}^\pi(V^{(t)}) - \mathcal{T}^\pi(V^\pi)\|_\infty \\
&\leq \gamma \|V^{(t)} - V^\pi\|_\infty \\
&\leq \gamma^t \|V^{(0)} - V^\pi\|_\infty.
\end{aligned}$$

Implication: to achieve $\|V^{(t+1)} - V^\pi\|_\infty \leq \epsilon$, it takes no more than

$$\frac{1}{1-\gamma} \log\left(\frac{\|V^{(0)} - V^\pi\|_\infty}{\epsilon}\right)$$

iterations.

Policy evaluation for finite-horizon MDPs



1. Begin with the terminal step $h = H + 1$:

$$V^\pi_{H+1} = 0, \quad Q^\pi_{H+1} = 0.$$

2. Backtrack $h = H, H-1, \ldots, 1$:

$$Q^\pi_h(s, a) := \underbrace{\mathbb{E}\left[r_h(s_h, a_h)\right]}_{\text{immediate reward}} + \underbrace{\mathbb{E}_{s' \sim P_h(\cdot|s,a)} V^\pi_{h+1}(s')}_{\text{next step's value}}$$

$$V^\pi_h(s) := \mathbb{E}_{a \sim \pi_h(\cdot|s)} Q^\pi_h(s, a)$$